Information Theory Considerations in Patch-based Training of Deep Neural Networks on Seismic Time-Series

Sören Dramsch, Jesper; Lüthje, Mikael

Published in:
Proceedings of the First EAGE/PESGB Workshop on Machine Learning (London2018)

Link to article, DOI:
10.3997/2214-4609.201803020

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Introduction

Sampling in physics-based applications and digital signal processing has long been recognised as an essential constraint. The Nyquist-Shannon theorem is the most prominent information theorem that prevents aliasing in seismic data (Seibt, 2006). Sampling has to be considered an essential part of a machine learning pipeline to avoid the implicit bias of learnt decision boundaries and joint distributions.

Machine Learning algorithms, particularly deep convolutional neural networks (CNN) often learn on patches of data. In many applications, the dynamic range of the data is additionally converted from 32-bit floats to 8-bit integers. This loss of dynamic range often speeds up training of networks and stabilises convergence at the loss of accuracy. However, investigations into precision have shown that this effect may be negligible (Holi and Hwang, 1993). Patch-based image training in machine learning usually takes smaller windows of data. The ImageNet challenge (Deng et al., 2009) provides 256x256 pixel images, which sets standards for many machine learning architectures.

Theory

Seismic traces are often sampled at 4ms and contain several hundred to thousands of samples. The Nyquist-Shannon theorem applies to high-frequency bounds only. However, we propose that a lower bound has to be adhered to when applying real-valued transformations to data before reconstruction. Low-frequency aliasing can be seen as a DC offset, where DC is the value at 0 Hz. This effect has been studied in non-stationary signals in applications such as seismic frequency decomposition (Chakraborty and Okaya 1995).

In statistical learning, many applications learn implicit joint distributions of the data. These are often approximated by multivariate distributions or transformations that operate solely on real-valued signals instead of complex signals (Hirose, 2003). This is equivalent to a mean shift of the data, as well as noise of the mean and may hinder convergence of the algorithms and diminish results. Inference on images that can appropriately sample low frequencies, due to a larger size, could lead to non-generalizability of the data due to implicit bias, which is the antithesis of machine learning.

We propose a low-frequency boundary, which follows the Nyquist-Shannon sampling theorem. With $f_{ny} = \frac{1}{2T}$, where T is the maximum period resolvable in the time series. This is due to the fact that we treat cutouts of a non-stationary signal as representative of the entire series and therefore, have to infer stationarity within the available bandwidth.

Example: Neural Network - Single Neuron

Neurons in neural networks are described by the activation $\sigma(w \cdot x + b)$, where w is the network weights, x is the input data, b is the network bias, and sigma is a non-linear activation function. A common non-linear activation is the rectified linear unit (RELU) $\sigma(x) = max(0, x)$. Considering the inference stage, the network weights w and biases b are fixed, x is the only variable parameter. Learning on a mean-shift of q of an arbitrary distribution over x leads to $\sigma(w \cdot (x + q) + b)$, which increases the neuron response by q, weighted by w. At inference, the mean-shift over larger inference data disappears, introducing an additional bias of $w \cdot q$ before non-linear activation. This training bias may lead to prediction errors of the neuron and consequently the full neural network.

Example: Dutch F3 Seismic data

We use a randomly selected trace from the Dutch F3 dataset. The total recording time is 4 seconds with 1001 samples sampled at 4 ms. The sampling interval of 4ms allows for a maximum frequency of 125 Hz. We compare the reconstruction of the signal from the real part of the frequency spectrum for non-overlapping patches. The frequency content of real-valued stationary traces would be similar, whether a trace is split into parts or whole.

The frequency content in Figure 1 shows that properly tapered data introduces a DC offset and the phase spectrum cannot be reconstructed fully. For a window of 101 samples at 4 ms, we get the lower Nyquist frequency of ~12.5 Hz. A patch of 256 samples at 4ms has the lower bound of ~5 Hz. We propose that a high-pass filter at training may improve convergence. Transfer learning on larger patches with fewer epochs then recovers low-frequency information, while keeping training times attainable.

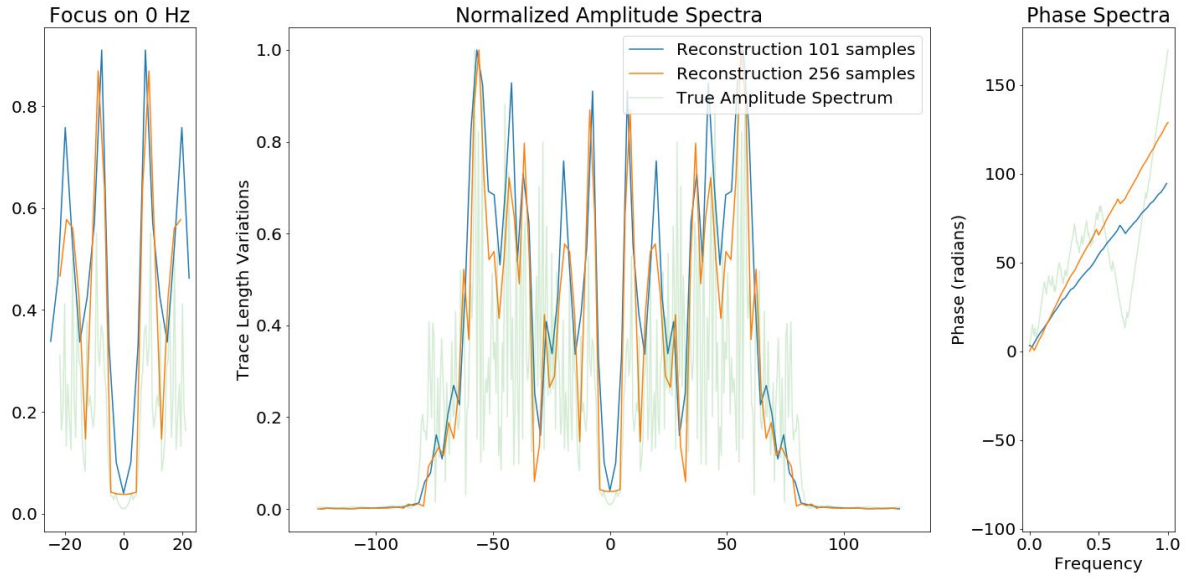

Figure 1 *We present different sizes of cutouts, with 101 and 256 samples respectively. In the middle, the full normalised amplitude spectra are presented. On the right, the according phase spectra are presented. On the left, we focus on the frequency content of the amplitude spectra around 0 Hz. The cutouts were Hanning tapered, however, a clear DC offset appears with decreasing patch size.*

Conclusions

We investigate the frequency content in non-overlapping patch-based seismic data. Non-overlapping patches may introduce low-frequency noise that translates to a mean-shift of learnt distributions. Further investigations into frequency responses of Convolutional Neural Networks (CNN) and the computation thereof, which is common in the frequency domain, should be undertaken. The authors note that signal processing paradigms apply to image-based CNNs and tapering of time-series before Fourier transformation is essential.

Acknowledgements

The research leading to these results has received funding from the Danish Hydrocarbon Research and Technology Centre under the Advanced Water Flooding program. The authors thank Matthias Schneider for fruitful discussion and dGB for providing the F3 dataset.

References

Avijit Chakraborty and David Okaya (1995). "Frequency-time decomposition of seismic data using wavelet-based methods." GEOPHYSICS, 60(6), 1906-1916

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). Ieee.

Hirose, A. (2003). Complex-Valued Neural Networks: An Introduction. In *Complex-Valued Neural Networks: Theories and Applications* (pp. 1-6).

Holi, J. L., & Hwang, J. N. (1993). Finite precision error analysis of neural network hardware implementations. *IEEE Transactions on Computers*, (3), 281-290.

Seibt, P. (2006). *Algorithmic Information Theory*. Springer.